ChatGPT and large language models: what's the risk?

Do loose prompts* sink ships? Exploring the cyber security issues of ChatGPT and LLMs.

David C, Paul J

Large language models (LLMs) and AI chatbots have captured the world's interest, ignited by the release of ChatGPT in late 2022 and the ease of querying it provides. It's now one of the fastest growing consumer applications ever, and its popularity is leading many competitors to develop their own services and models, or to rapidly deploy those that they've been developing internally.

As with any emerging technology, there's always concern around what this means for security. This blog considers some cyber security aspects of ChatGPT and LLMs more generally in the near term.

What is ChatGPT and what are LLMs?

ChatGPT is an artificial intelligence chatbot developed by OpenAI, a US tech startup. It's based on GPT-3, a language model released in 2020 that uses deep learning to produce human-like text, but the underlying LLM technology has been around much longer.

An LLM is where an algorithm has been trained on a large amount of text-based data, typically scraped from the open internet, and so covers web pages and - depending on the LLM - other sources such as scientific research, books or social media posts. This covers such a large volume of data that it's not possible to filter all offensive or inaccurate content at ingest, and so 'controversial' content is likely to be included in its model.

The algorithms analyse the relationships between different words and turn that into a probability model. It is then possible to give the algorithm a 'prompt' (for

example, by asking it a question), and it will provide an answer based on the relationships of the words in its model.

Typically, the data in its model is static after it has been trained, although it can be refined by 'fine-tuning' (which is training on additional data) and 'prompt augmentation' (which is providing context information about the question). An example of prompt augmentation might be:

Taking into account the below information, how would you describe...

and then copying potentially large amounts of text (or whole documents) into the prompt/question.

ChatGPT effectively allows users to ask an LLM questions, as you would when holding a conversation with a chatbot. Other recent examples of LLMs include the announcement of Google's Bard and Meta's LLaMa (for scientific papers).

LLMs are undoubtedly impressive for their ability to generate a huge range of convincing content in multiple human and computer languages. However, they're not magic, they're not artificial general intelligence, and contain some serious flaws, including:

- they can get things wrong and 'hallucinate' incorrect facts
- they can be biased, are often gullible (in responding to leading questions, for example)
- they require huge compute resources and vast data to train from scratch
- they can be coaxed into creating toxic content and are prone to 'injection attacks'

Will LLMs reveal my information?

A common concern is that an LLM might 'learn' from your prompts, and offer that information to others who query for related things. There is some cause for concern here, but not for the reason many consider. Currently, LLMs are trained, and then the resulting model is queried. An LLM does **not** (as of writing)

automatically add information from queries to its model for others to query. That is, including information in a query will not result in that data being incorporated into the LLM.

However, the query **will** be visible to the organisation providing the LLM (so in the case of ChatGPT, to OpenAI). Those queries are stored and will almost certainly be used for developing the LLM service or model at some point. This could mean that the LLM provider (or its partners/contractors) are able to read queries, and may incorporate them in some way into future versions. As such, the **terms of use** and **privacy policy** need to be thoroughly understood before asking sensitive questions.

A question might be sensitive because of data included in the query, or because who is asking the question (and when). Examples of the latter might be if a CEO is discovered to have asked 'how best to lay off an employee?', or somebody asking revealing health or relationship questions. Also bear in mind aggregation of information across multiple queries using the same login.

Another risk, which increases as more organisations produce LLMs, is that queries stored online may be hacked, leaked, or more likely accidentally made publicly accessible. This could include potentially user-identifiable information. A further risk is that the operator of the LLM is later acquired by an organisation with a different approach to privacy than was true when data was entered by users.

As such, the NCSC recommends:

- not to include sensitive information in queries to public LLMs
- not to submit queries to public LLMs that would lead to issues were they made public

---

How can I safely provide LLMs with sensitive information?

In the wake of the excitement around LLMs, many organisations may be wondering if they can use LLMs to automate certain business tasks, which may

involve providing sensitive information either through fine-tuning or prompt augmentation. Whilst this approach is **not** recommended for public LLMs, 'private LLMs' might be offered by a **cloud provider** (for example), or can be entirely **self hosted**:

- For **cloud-provided LLMs**, the terms of use and privacy policy again become key (as they are for public LLMs), but are more likely to fit within the existing terms for the cloud service. Organisations need to understand how the data they use for fine-tuning or prompt augmentation is managed. Is it available to the vendor's researchers or partners? If so, in what form? Is data shared in isolation or in aggregation with other organisations? Under what conditions can an employee at the provider view queries?
- **Self-hosted LLMs** are likely to be highly expensive. However, following a security assessment (which should include referring to the NCSC's principles for the security of machine learning), they may be appropriate for handling organisational data. In particular, organisations should refer to our guidance on securing your infrastructure and data supply chains.

Do LLMs make life easier for cyber criminals?

There have been some incredible demonstrations of how LLMs can help write malware. The concern is that an LLM might help someone with malicious intent (but insufficient skills) to create tools they would not otherwise be able to deploy. In their current state, LLMs suffer from *appearing* convincing (whether or not they are), and are suited to simple tasks rather than complex ones. This means LLMs are useful for 'helping experts save time', as the expert can validate the LLM's output.

For more complex tasks, it's currently easier for an expert to create the malware from scratch, rather than having to spend time correcting what the LLM has produced. However, an expert capable of creating highly capable malware is likely to be able to coax an LLM into writing capable malware. This trade-off between *'using LLMs to create malware from scratch'* and *'validating malware created by LLMs'* will change as LLMs improve.

LLMs can also be queried to advise on technical problems. There is a risk that criminals might use LLMs to help with cyber attacks beyond their current capabilities, in particular once an attacker has accessed a network. For example, if an attacker is struggling to escalate privileges or find data, they might ask an LLM, and receive an answer that's not unlike a search engine result, but with more context. Current LLMs provide *convincing-sounding* answers that may only be partially correct, particularly as the topic gets more niche. These answers might help criminals with attacks they couldn't otherwise execute, or they might suggest actions that hasten the detection of the criminal. Either way, the attacker's queries will likely be stored and retained by LLM operators.

As LLMs excel at replicating writing styles on demand, there is a risk of criminals using LLMs to write convincing phishing emails, including emails in multiple languages. This may aid attackers with high technical capabilities but who lack linguistic skills, by helping them to create convincing phishing emails (or conduct social engineering) in the native language of their targets.

To summarise, in the near term we might see:

- more convincing phishing emails as a result of LLMs
- attackers trying techniques they didn't have familiarity with previously

There is also a low risk of a lesser-skilled attacker writing highly capable malware.

To wrap up

It's an exciting time for LLMs, and ChatGPT in particular has gripped the world's imagination. As with all technology developments, there will be people keen to use it and to investigate what it has to offer, and those who may never use it.

There are undoubtedly risks involved in the unfettered use of public LLMs, as we've outlined above. Individuals and organisations should take great care with the data they choose to submit in prompts. You should ensure that those who want to experiment with LLMs are able to, but in a way that doesn't place organisational data at risk.

The NCSC is aware of other emerging threats (and opportunities) in relation to cyber security and adoption of LLMs, and we will of course keep you informed of these in future blogposts.

David C - Tech Director for Platforms Research
Paul J - Tech Director for Data Science Research

* *Loose lips sink ships* was the US equivalent of the UK's *Careless talk costs lives.* The United States Office of War Information created and used it during World War II on posters to encourage people to avoid careless talk.

WRITTEN BY

David C
NCSC Technical Director for Platforms Research

WRITTEN BY

Paul J
Technical Director for Data Science Research, NCSC

PUBLISHED

14 March 2023

WRITTEN FOR

Cyber security professionals

Public sector

Large organisations